# Crosswalk Recognition Through Point-Cloud Processing and Deep-Learning Suited to a Wearable Mobility Aid for the Visually Impaired

Matteo Poggi(✉), Luca Nanni, and Stefano Mattoccia

Department of Computer Science and Engineering (DISI), University of Bologna, Viale Risorgimento, 2, 40136 Bologna, Italy
{matteo.poggi8,stefano.mattoccia}@unibo.it, luca.nanni10@studio.unibo.it

**Abstract.** In smart-cities, computer vision has the potential to dramatically improve the quality of life of people suffering of visual impairments. In this field, we have been working on a wearable mobility aid aimed at detecting in real-time obstacles in front of a visually impaired. Our approach relies on a custom RGBD camera, with FPGA on-board processing, worn as traditional eyeglasses and effective point-cloud processing implemented on a compact and lightweight embedded computer. This latter device also provides feedback to the user by means of an haptic interface as well as audio messages. In this paper we address crosswalk recognition that, as pointed out by several visually impaired users involved in the evaluation of our system, is a crucial requirement in the design of an effective mobility aid. Specifically, we propose a reliable methodology to detect and categorize crosswalks by leveraging on point-cloud processing and deep-learning techniques. The experimental results reported, on 10000+ frames, confirm that the proposed approach is invariant to head/camera pose and extremely effective even when dealing with large occlusions typically found in urban environments.

**Keywords:** Wearable · Embedded 3d vision · Deep learning · Crosswalk detection

## 1 Introduction and Related Work

Autonomous mobility, especially in urban environments, can be a challenging task for people suffering of visual impairments. Although some stationary obstacles can be learned day by day, many others change dynamically and thus can't be learned. For this reason, several mobility devices aimed at detecting obstacles, possibly by means of a contact-less strategy, have been proposed. Nevertheless, despite this fact, this strategy is not adopted by the white cane, the most widely adopted mobility aid by visually impaired users. Moreover, the white cane does not allow to perceive other crucial features such as pedestrian crossings.

Many vision-based systems have been proposed to deal with crosswalk recognition, or more in general urban road markings recognition, for different purposes. Most devices were proposed for vehicles, as assistive device as well as as


V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 282–289, 2015.
DOI: 10.1007/978-3-319-23222-5_35

part of autonomous driving systems, such as [1] that detects crosswalks by applying several filters on 2D images and [2] that relies on a bird-view re-projection of the 2D image. Some methods exploit 3D data [3,4] while others also rely on non-vision techniques; for example, Suzuki et al. [5] use 2D image processing and radar technology. In this field, an interesting study, aimed at analyzing drivers behavior in presence of different urban road markings, has been proposed in [6].

Other approaches have been designed to aid the visually impaired. In [7], an effective methodology was proposed to detect crosswalks, estimating their extension, and traffic lights, detecting the emitted color. Some of them, such as [8] and [9], have been implemented on a smartphone. Radvanyi et al. [10] proposed a wearable device based on a neural network to detect ground plane in 2D images and then recognizing crosswalks. In [11], the 3D data obtained through a stereo vision system is processed applying the Hough transform in the 2D and 3D domain to detect crosswalks and stairs. Crosswatch system [12] allows self localization by recognizing specific street patterns. In this paper we propose an effective methodology to detect crosswalks by leveraging 3D data provided by a custom RGBD camera and a *Convolutional Neural Network* (CNN).

## 2 Overview of the Wearable Mobility Aid

In this section we provide a brief overview of our wearable mobility aid for obstacle detection, proposed in its early development phase in [13].

It consists of a custom RGBD sensor developed by our research group [14], based on stereo vision technology, and an embedded ARM board. Our system is purely based on vision technology and is powered by a small accumulator that enables hours of battery life. The 3D sensor provides dense and accurate depth map processing synchronized stereo images at more than 30 fps (up to 640 $\times$ 480 resolution) according to state-of-the-art 3D vision algorithms implemented into a low cost FPGA (Spartan 6 model 75 in the current setup). Specifically, we have mapped into the FPGA a complete stereo vision pipeline including a custom and modified version of the SGM algorithm [15]. The output of the RGBD sensor (reference rectified image and disparity map as shown in Figure1 a) and b)) is sent, via USB at about 20 fps, to the embedded computer, Odroid U3 [16], for obstacle detection. The early stage of the visual processing pipeline, greatly improved wrt the implementation shown in [13] consists of the following steps: disparity map to point-cloud conversion, ground plane segmentation according to a robust RANSAC [17] framework applied to the point-cloud, head pose estimation wrt the ground plane and refinement based on Kalman filtering. Once obtained the ground plane equation and the head pose we re-project, from a top-view perspective, points not laying on the ground plane and in this domain we compute, within vertical bins (of size 2 $\times$ 2 cm in the current setup), statistics concerned with heights (e.g., min and max values) and occupancy to accurately detect potential obstacles. According to suggestions provided by visually impaired users involved in the testing phase, the original field of view of the camera is restricted to the three nearby VOIs shown in Figure 1 c). Tactile and

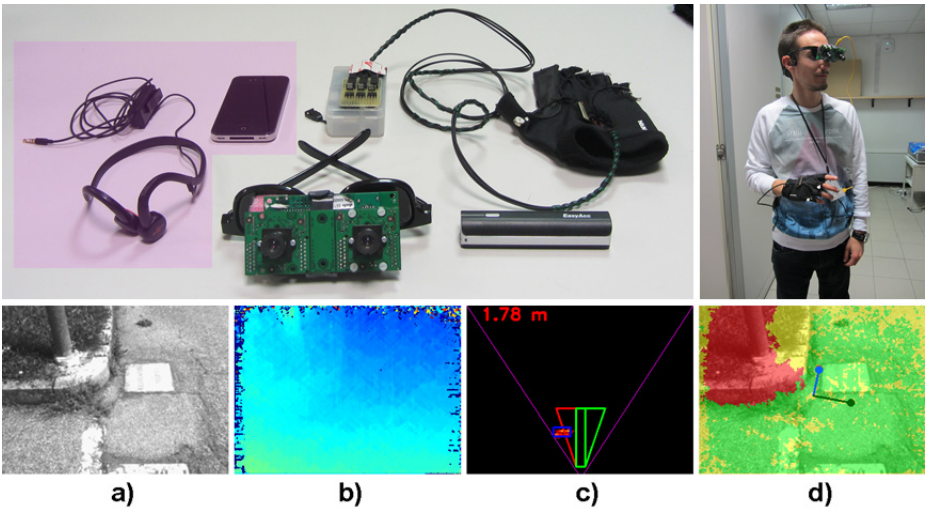


**Fig. 1.** On top, the adopted wearable mobility aid. It consists of a custom RGBD sensor, an Odroid U3 system, a haptic glove, a battery (enabling 3+ hours autonomy) and optional audio interfaces (purple). On bottom, overview of the obstacle detection approach deployed on it. a) Reference rectified image - b) Disparity map computed on FPGA (colder colors encode farther points) - c) Top-view re-projection with three sensed *volumes of interest* (VOI) in front of the user and, highlighted, the obstacles - d) Segmented ground plane (green) with superimposed head pose wrt the ground plane and detected obstacle regions (red).

audio feedback are provided by means of a vibro-tactile glove, bone conductive headset and smartphone. The whole hardware setup described so far and depicted in Figure 1 weights about 250 g, including the battery.

## 3 Proposed Crosswalk Recognition Approach

In this paper we propose a crucial enhancement to the outlined mobility aid providing reliable crosswalk recognition. This additional feature, not available with a white cane, would greatly improve the knowledge of the explored environment enabling the visually impaired to properly locate the presence and the direction of pedestrian crossings as well as to improve his/her self localization. To detect crosswalk and recognize their orientation, according to the four categories depicted at the left of Figure 2, wrt the user we rely on point-cloud processing and a CNN. In particular, two main phases are carried out:

- Head pose estimation, aimed at warping the ground plane according to the estimated head position computed from point-cloud data provided by the RGBD sensor
- Detection of pedestrian crossings and categorization of their orientation with respect to the user on the segmented and warped ground plane images

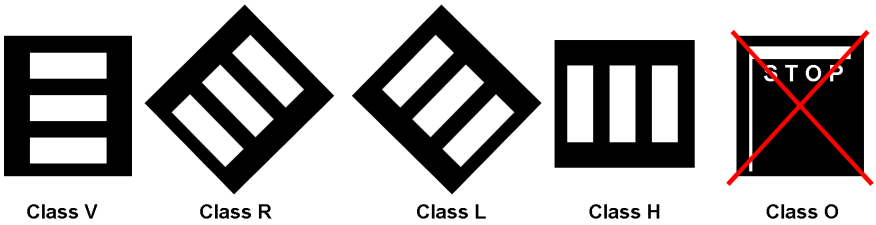


**Fig. 2.** Proposed ontology. The CNN is trained to detect and classify pedestrian crossings according to four possible orientation wrt the user's point of view: *vertical* (V), *horizontal* (H), *diagonal left* (L) and *diagonal right* (R). A further class, referred to as *other* (O), takes into account any other case.

### 3.1 Head Pose Estimation and Image Refinement

Our system, starting from the dense disparity map provided by the RGBD sensor, computes on the embedded CPU the point-cloud according to (1) mapping each point with a valid disparity value to the corresponding 3D point of coordinates $(X_c, Y_c, Z_c)$ wrt the camera reference system by knowing the baseline of the stereo camera $b$, the focal length $f$, the optical center $(u_0, v_0)$ and the coordinate $(u, v)$ of the point at disparity $d$.

$$Z_c = \frac{bf}{d} \qquad X_c = \frac{Z_c(u - u_0)}{f} \qquad Y_c = \frac{Z_c(v - v_0)}{f} \tag{1}$$

From the point-cloud, a robust RANSAC framework [17] allows us to obtain a reliable estimation of the ground plane equation. This information enables to discriminate between ground plane (where crosswalk markers are painted) and any other object not laying on this surface. Then, the segmented ground plane image/point-cloud is further processed before getting analyzed by the CNN, which could wrongly estimate the direction taken by the crosswalk in presence of head/camera tilting. In particular, we found that recognition accuracy improves when the head/camera is aligned to the floor (i.e., when the normal vector of the ground plane, if drawn on the image, appears to be vertical). To follow this strategy, the angle that aligns the ground plane normal with the vertical direction is computed and used to warp the segmented image accordingly. Once obtained such normalized representation of the ground plane from point-cloud data, the warped image can be processed by the CNN to detect and classify potential pedestrian crossings.

### 3.2 CNN Architecture for Crosswalk Recognition

Machine learning techniques have been widely adopted in many practical applications and deep-learning is one of the most effective techniques for visual recognition. A deep neural network is a multilayer architecture with layers connected by non-linear transformations. In computer vision, CNNs are deep neural networks made of several layers, called *convolutional layers*, that extract features

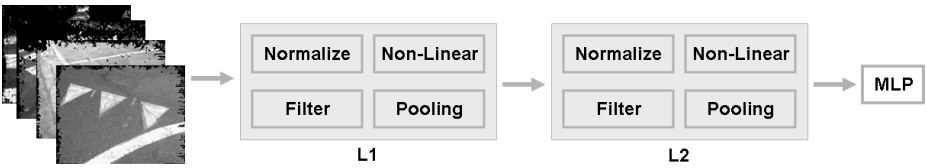


**Fig. 3.** Architecture of the CNN based on two convolutional layers showing their modules.

from the images by applying several normalization and filtering operations, and a final classifier, typically, a *Multi Layer Perceptron* (MLP). Compared to other machine learning techniques, such as Bag of Visual Words [18], that rely on an explicit feature extraction phase, a CNN allows for a higher level of abstraction deploying adaptive convolutional layers. LeCun et al. [19] reported how such multistage architectures yield to significant improvements compared to a single layer architecture.

In our approach the CNN takes as input the warped image of the ground plane, detects the presence of a crosswalk and, if found, its orientation. The user is made aware of the outcome of this process with an audio message. In our architecture, we adopt a 2-layers plus MLP structure, as shown in Figure 3, mapped within the Torch 7 framework [20]. Specifically, the two convolutional layers and the MLP have been designed as follows:

- Layer 1: *Filter* performs spatial convolution to extracts 256 10×10 feature maps by using 5×5 filters and fan-in equal to 1, *Non-Linear* applies hyperbolic tangent as squashing function (enhancing strong features and suppressing weak ones), *Pooling*, on 2×2 regions and 2×2 stride, obtaining 16 14×14 maps, *Normalize* performs feature normalization
- Layer 2: *Filter* performs spatial convolution to extracts 16 28×28 feature maps by using 5×5 filters and fan-in equal to 4, *Non-Linear* applies hyperbolic tangent as in Layer 1, *Pooling*, on 2×2 regions and 2×2 stride, obtaining 256 5×5 maps, *Normalize* performs feature normalization
- *MLP*: made by a 128 neurons level fully connected to a 5 neurons further level, adopting hyperbolic tangent for back propagation

## 4 Experimental Results

For the experimental validation we trained the CNN on a dataset composed of about 2500 images acquired with our wearable mobility aid in urban scenarios. For each of the 5 classes depicted in Figure 2 we have acquired about 500 training instances. After a 15 *epochs* training period the test set composed of 100 images per class has been subject of categorization returning a 100% correctness ratio. Eventually, to properly evaluate the effectiveness of the proposed approach in challenging urban environments including scenes with large and multiple occlusions, difficult illumination conditions, ruined crosswalk patterns and so on, we

**Table 1.** Confusion matrices computed on the validation set (10165 frames). On the left, by processing the raw segmented images, we obtain 0 false negatives (i.e., undetected crosswalks) and 741 false positives (5.97%). On the right, by processing the segmented images after head pose refinement, we obtain 0 false negatives and 995 false positives (6.63%).

| | | | | | |
|---|---|---|---|---|---|
| 1983 | 162 | 96 | 0 | 0 | V, 88.48% |
| 58 | 814 | 0 | 19 | 0 | R, 91.35% |
| 45 | 0 | 874 | 14 | 0 | L, 93.67% |
| 0 | 4 | 3 | 97 | 0 | H, 93.27% |
| 60 | 58 | 78 | 411 | 5369 | O, 89.84% |

| | | | | | |
|---|---|---|---|---|---|
| 2198 | 19 | 24 | 0 | 0 | V, **98.08%** |
| 28 | 859 | 0 | 4 | 0 | R, **96.40%** |
| 25 | 0 | 903 | 5 | 0 | L, **96.78%** |
| 0 | 0 | 3 | 101 | 0 | H, **97.12%** |
| 95 | 65 | 91 | 423 | 5302 | O, 88.72% |

acquired a validation dataset composed of 10000+ frames. In Table 1 we report *confusion matrices* summarizing the results on such dataset. A confusion matrix has N rows and N columns, with N the number of classes in the ontology, and highlights the following:

- The main diagonal contains the number of correct instances for each class
- On each row, the accuracy percentage is reported for a class, showing how many frames are miss-recognized and the class they are wrongly assigned to
- On each column, it shows how many frames, for each other class, are wrongly categorized as belonging to a different class

The table, on the left, shows results concerning recognition accuracy obtained by processing the raw ground plane segmentation image without head pose refinement. We can notice a high correctness rate for crosswalk recognition, which is between 88% and 94% for each of the four classes V, R, L, H. Moreover, it is worth noticing that wrongly categorized crosswalk frames are always assigned to a different zebra crossing pattern and never miss-categorized as class O. Therefore, the crosswalk recognition rate is 100%. On the other hand, we can also notice that we have 10.16% of the frames belonging to class O wrongly classified, resulting in a false positive percentage (i.e., images wrongly categorized as crosswalk) of 5.97% on the overall validation set.

Applying the head pose refinement, on the right in Table 1, we obtain an average recognition accuracy improvement between 3% and 5% on classes R, L and H, with a major increase close to 10% for V. Figure 4 shows 2 out 10000+ frames of the validation dataset; in particular the first row reports a scenario where the head pose refinement phase allows to detect the correct class (H). On the other hand, the number of frames of class O wrongly categorized as crosswalk slightly increases by less than 1% on the overall dataset.

In general, false positives are mainly due to particular challenging environments when framing regions containing shadows close to areas exposed to sunlight. However, it is worth observing that our current training set has a limited cardinality and an extended dataset, currently under acquisition, would certainly allow to further reduce the number of false detections. Finally, we report that on the Odroid U3 our approach computes plane detection plus head pose

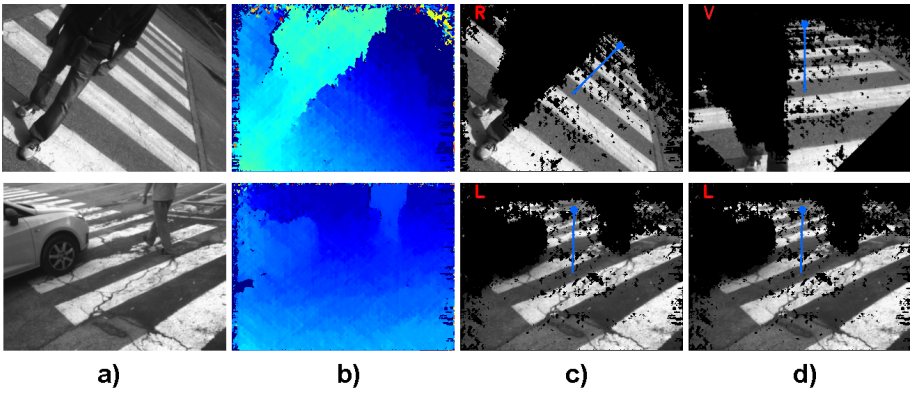


**Fig. 4.** Two frames from the validation set (10000+ instances). a) the reference rectified image - b) the raw disparity map computed by the RGBD sensor - c) the detected ground plane - d) refined/warped ground plane according to the normal vector. c) and d) also show the recognized orientation, corrected by pose refinement on the first frame.

refinement in about 20 ms and crosswalk recognition in 180 ms thus allowing a prompt feedback to the user.

## 5 Conclusions

In this paper, an effective crosswalk recognition pipeline leveraging 3D data provided by a compact RGBD sensor and deep-learning has been proposed. Despite the small cardinality of the current training set, experimental results on 10000+ images acquire in challenging urban environments, show a quite high recognition accuracy even in presence of large occlusions. Moreover, its computational efficiency makes it suitable for real-time crosswalk recognition even on the target embedded device deployed for a wearable mobility aid. A larger dataset would improve the accuracy, with no computational overhead and this approach could also be extended to detect and recognize other road markings for several purposes.

## References

1. Haselhoff, A., Kummert, A.: On visual crosswalk detection for driver assistance systems. In: 2010 IEEE Intelligent Vehicles Symposium (IV), pp. 883–888, June 2010
2. Wu, T., Ranganathan, A.: A practical system for road marking detection and recognition. In: 2012 IEEE Intelligent Vehicles Symposium (IV), pp. 25–30, June 2012
3. Mancini, A., Frontoni, E., Zingaretti, P.: Automatic road object extraction from mobile mapping systems. In: 2012 IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications (MESA), pp. 281–286, July 2012

4. Hata, A., Wolf, D.: Road marking detection using lidar reflective intensity data and its application to vehicle localization. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), pp. 584–589, October 2014
5. Suzuki, S., Raksincharoensak, P., Shimizu, I., Nagai, M., Adomat, R.: Sensor fusion-based pedestrian collision warning system with crosswalk detection. In: 2010 IEEE Intelligent Vehicles Symposium (IV), pp. 355–360, June 2010
6. Ishizaki, R., Morimoto, M., Fujii, K.: An evaluation method of driving behavior by in-vehicle data camera. In: 2012 Fifth International Conference on Emerging Trends in Engineering and Technology (ICETET), pp. 293–297, November 2012
7. Shioyama, T., Wu, H., Nishibe, Y., Nakamura, N., Kitawaki, S.: Image analysis of crosswalk. In: proceedings of the 11th International Conference on Image Analysis and Processing, pp. 168–173, September 2001
8. Ivanchenko, V., Coughlan, J., Shen, H.: Detecting and locating crosswalks using a camera phone. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2008, pp. 1–8, June 2008
9. Ahmetovic, D., Bernareggi, C., Gerino, A., Mascetti, S.: Zebrarecognizer: efficient and precise localization of pedestrian crossings. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 2566–2571, August 2014
10. Radvanyi, M., Varga, B., Karacs, K.: Advanced crosswalk detection for the bionic eyeglass. In: 2010 12th International Workshop on Cellular Nanoscale Networks and Their Applications (CNNA), pp. 1–5, February 2010
11. Wang, S., Tian, Y.: Detecting stairs and pedestrian crosswalks for the blind by rgbd camera. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 732–739, October 2012
12. Murali, V.N., Coughlan, J.M.: Smartphone-based crosswalk detection and localization for visually impaired pedestrians. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–7, July 2013
13. Mattoccia, S., Macri', P.: 3d glasses to improve autonomous mobility of people visually impaired. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV Workshop. LNCS, vol. 8927, pp. 539–554. Springer, Switzerland (2014)
14. Mattoccia, S., Marchio, I., Casadio, M.: A compact 3d camera suited for mobile and embedded vision applications. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 195–196, June 2014
15. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)
16. Hard-Kernel: Odroid u3. http://hardkernel.com/main/main.php
17. Choi, S., Kim, T., Yu, W.: Performance evaluation of ransac family. In: BMVC (2009)
18. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
19. Jarrett, K., Kavukcuoglu, K., Ranzato, M.A., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: Proc. International Conference on Computer Vision (ICCV 2009). IEEE (2009)
20. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: a matlab-like environment for machine learning. In: BigLearn, NIPS Workshop (2011)